Perturbation Theory

Suppose $A\vec{x} = \vec{b}$ and we approximately solve by a backwards stable algorithm, obtaining $\hat{x}$ so that

$$(A + \delta A)\hat{x} = b + \delta b$$

we'd like to estimate $\delta x = \hat{x} - x$, which is the error in the solution we computed. We write

$$\begin{array}{rl} (A+\delta A)(x+\delta x) &= b + \delta b \\ - \qquad\qquad Ax &= b \\ \hline \delta A\, x + (A+\delta A)\delta x &= \delta b \end{array}$$

or

$$(A + \delta A)\,\delta x = \delta b - \delta A\, x$$

$$\begin{aligned} A\,\delta x &= \delta b - \delta A\, x - \delta A\,\delta x \\ &= \delta b - \delta A\,\hat{x} \end{aligned}$$

b.

or

$$\delta x = A^{-1}(\delta b - \delta A \hat{x}).$$

Using our norm properties,

$$\begin{aligned}
\|\delta x\| &= \|A^{-1}(\delta b - \delta A \hat{x})\| \\
&\leq \|A^{-1}\| \, \|\delta b - \delta A \hat{x}\| \quad \text{triangle inequality} \\
&\leq \|A^{-1}\| \left(\|\delta b\| + \|\delta A \hat{x}\| \right) \\
&\leq \|A^{-1}\| \left(\|\delta b\| + \|\delta A\| \, \|\hat{x}\| \right).
\end{aligned}$$

We can write this in terms of the relative error in the solution by rearranging:

$$\underbrace{\frac{\|\delta x\|}{\|\hat{x}\|}}_{\text{error in answer}} \leq \|A^{-1}\| \, \|A\| \underbrace{\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \, \|\hat{x}\|} \right)}_{\text{error in input data A and b}}$$

We call

$$K(A) = \|A^{-1}\| \, \|A\|$$

the condition number of A (as a matrix) because it determines the sensitivity ~~to~~ of the solution to $A\vec{x}=b$ to the input data A and b.

Of course, in practice, $\hat{x}$ is what you <u>have</u>, so the bound

$$\frac{\|\delta x\|}{\|\hat{x}\|} \le K(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \, \|\hat{x}\|}\right)$$

is perfectly usable. But it's not very good for developing a theory because we'd like to express everything in terms of the <u>actual</u> solution ~~s~~ ~~s~~ x.

d.

We can do this with a little more work.

Lemma. Suppose $\| \ \|$ is a matrix norm which is self-consistent (so $\|AB\| \leq \|A\| \, \|B\|$). ~~Then~~ If $\|X\| < 1$, then $I - X$ is invertible

$$(I-X)^{-1} = \sum_{i=0}^{\infty} X^i$$

and

$$\|(I-X)^{-1}\| \leq \frac{1}{1-\|X\|}$$

Proof. The infinite series of matrices $\sum X^i$ converges $\Leftrightarrow$ it converges componentwise. Now for any norm, $\exists$ some constant c so that

$$|X_{jk}| \leq c \, \|X\|$$

for any matrix X (of given dimensions).

So

$$\underbrace{|(X^i)_{jk}|}_{\substack{jk\text{-th entry}\\ \text{of } X^i}} \leq c\|X^i\| \leq c\|X\|^i$$

But $\|X\| < 1$, so this thing at right is a convergent geometric series

$$\sum_{i=0}^{\infty} c\|X\|^i = \frac{c}{1-\|X\|}$$

and the series

$$\sum |(X^i)_{jk}|$$

must converge as well. Thus the sequence of partial sums $S_n = \sum_{i=0}^{n} X^i$ converges to some matrix S.

Now we play the polynomial game:

$$(I-X)S_n = (I+X+X^2+\dots+X^n)(I-X) = I - X^{n+1}$$

and we see $I - X^{n+1} \to I$ as $n \to \infty$ since $\| X^{n+1} \| \leq \| X \|^{n+1}$, which is going to zero.

So

$$(I - X) S = I$$

and

$$S = (I - X)^{-1},$$

proving the first part of the claim.

To see the norm bound, we write

$$\| (I-X)^{-1} \| = \| \Sigma X^i \| \leq \Sigma \| X^i \|$$

$$\leq \Sigma \| X \|^i = \frac{1}{1 - \| X \|}. \quad \square$$

Now we can eliminate the ~~δx~~ $\hat{X}$ s from our previous bound.

Recall that

$$(A+\delta A)(x+\delta x) = b+\delta b$$

so since $Ax=b$, we have

$$\delta A\, x + (A+\delta A)\delta x = \delta b$$

and

$$\begin{aligned}\delta x &= (A+\delta A)^{-1}(\delta b - \delta A x).\\ &= \left[A(I+A^{-1}\delta A)\right]^{-1}(\delta b - \delta A x)\\ &= (I+A^{-1}\delta A)^{-1}A^{-1}(\delta b - \delta A x).\end{aligned}$$

Taking norms and dividing by $\|x\|$,

$$\frac{\|\delta x\|}{\|x\|} \leq \|(I+A^{-1}\delta A)^{-1}\|\,\|A^{-1}\|\left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|}\right)$$

Now

$$\|A^{-1}\delta A\| \leq \|A^{-1}\|\,\|\delta A\|,$$

and if $\|\delta A\|$ is small enough, this is < 1. This means that our Lemma applies and

$$\|(I+A^{-1}\delta A)^{-1}\| \leq \frac{1}{1-\|A^{-1}\delta A\|} \leq \frac{1}{1-\|A^{-1}\|\,\|\delta A\|}$$

So we have

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\,\|\delta A\|} \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right)$$

$$\leq \frac{\|A^{-1}\|\,\|A\|}{1 - \|A^{-1}\|\,\|A\| \left(\frac{\|\delta A\|}{\|A\|} \right)} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\|\,\|x\|} \right)$$

but $\|A\|\,\|x\| \geq \|Ax\| = \|b\|$, so we can replace the denominator of the last fraction at right and get

$$\leq \frac{K(A)}{1 - K(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right).$$

Now if the abs ~~relative~~ error $\|\delta A\|$ is small, the quantity $1 - K(A) \frac{\|\delta A\|}{\|A\|} = 1 - \|A^{-1}\|\,\|\delta A\|$ is close to 1 and this is close to

$$\approx\leq K(A) \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right).$$